Employing Bilinear Fusion and Saliency Prior Information for RGB-D Salient Object Detection

Huang, Nianchang; Yang, Yang; Zhang, Dingwen; Zhang, Qiang; Han, Jungong

Published in:
IEEE Transactions on Multimedia

DOI:
10.1109/TMM.2021.3069297

Publication date:
2022

Citation for published version (APA):
Huang, N., Yang, Y., Zhang, D., Zhang, Q., & Han, J. (2022). Employing Bilinear Fusion and Saliency Prior Information for RGB-D Salient Object Detection. *IEEE Transactions on Multimedia*, *24*, 1651-1664. https://doi.org/10.1109/TMM.2021.3069297

Employing Bilinear Fusion and Saliency Prior Information for RGB-D Salient Object Detection

Nianchang Huang, Yang Yang, Dingwen Zhang, Qiang Zhang*, Jungong Han*

***Abstract*—Multi-modal feature fusion and saliency reasoning are two core sub-tasks of RGB-D salient object detection. However, most existing models employ linear fusion strategies (e.g., concatenation) for multi-modal feature fusion and use a simple coarse-to-fine structure for saliency reasoning. Despite their simpleness, they can neither fully capture the cross-modal complementary information nor exploit the multi-level complementary information among the cross-modal features at different levels. To address these issues, a novel RGB-D salient object detection model is presented, where we pay special attention to the aforementioned two sub-tasks. Concretely, a multi-modal feature interaction module is first presented to explore more interactions between the unimodal RGB and depth features. It helps to capture their cross-modal complementary information by jointly using some simple linear fusion strategies and bilinear fusion ones. Then, a saliency prior information guided fusion module is presented to exploit the multi-level complementary information among the fused cross-modal features at different levels. Instead of employing a simple convolutional layer for the final saliency prediction, a saliency refinement and prediction module is designed to better exploit those extracted multi-level cross-modal information for RGB-D saliency detection. Experimental results on several benchmark datasets verify the effectiveness and superiority of the proposed framework over some state-of-the-art methods.**

***Index Terms*—RGB-D salient object detection, bilinear fusion strategy, saliency prior information guided fusion, saliency refinement and prediction.**

I. Introduction

SALIENT object detection (SOD) [1], [2], [3], [4], [5], [6], [7] is a fundamental yet challenging task in computer vision. It seeks to detect the most visually distinctive regions in a given image. SOD plays an important role in many computer vision tasks, such as image classification [8], visual tracking [9] and segmentation [10]. Benefiting from the progress of Convolutional Neural Networks (CNNs), CNNs based RGB SOD models [2], [11], [12], [13] have significantly improved the performance of conventional hand-crafted feature based approaches [14], [15], [16], [17].

However, such algorithms are found vulnerable to complex environments, varying illuminations or cluttered backgrounds. After paying a lot of efforts, researchers realize that using RGB images only cannot solve those challenges. Meanwhile, the depth image seems to provide some geometrical information about the scene, which is invariant to the changes of illuminations and cluttered backgrounds. Furthermore, paired RGB and depth (RGB-D) images are easily captured by using some advanced depth cameras [18], [19]. Therefore, RGB-D SOD has seen an ever-increasing interest recently. So far, many RGB-D SOD methods have been proposed to exploit the cross-modal complementary information in RGB-D images for saliency detection [4], [5], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29].

(a) (b) (c) (d) (e) (f) (g)

Fig. 1. Visualization of the cross-modal features obtained by using different fusion strategies. (a) RGB images; (b) Depth images; (c) Cross-modal features obtained by employing concatenation operation; (d) Cross-modal features obtained by the proposed MFI module; (e) and (f) Saliency maps deduced from the cross-modal features obtained by using concatenation and the proposed fusion module, respectively; (g) Ground truth.

Generally, an RGB-D SOD model consists of three sub-models: unimodal feature extraction, multi-modal feature fusion and saliency reasoning. Unimodal feature extraction focuses on how to extract discriminative unimodal features from the input RGB-D images. Multi-modal feature fusion focuses on how to effectively fuse the unimodal (RGB and depth) features for better capturing the cross-modal complementary information. Saliency reasoning focuses on how to fully exploit the extracted cross-modal complementary information for saliency detection. Considering that unimodal feature extraction has been well studied in RGB SOD tasks [2], [11], [12], [13], we mainly investigate the problems of multi-modal feature fusion and saliency reasoning in this paper.

(1) How to effectively fuse unimodal RGB and depth features for better capturing the cross-modal complementary information.

Given the unimodal RGB and depth features, most existing RGB-D SOD models employ linear fusion strategies (e.g., concatenation or element-wise addition) to fuse multi-modality features [4], [5], [20], [21]. This may work well in most

Nianchang Huang, Yang Yang, Dingwen Zhang, Qiang Zhang are with Center for Complex Systems, School of Mechano-Electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, China. Email: nchuang@stu.xidian.edu.cn, yang@stu.xidian.edu.cn, zdw@xidian.edu.cn and qzhang@xidian.edu.cn.

Jungong Han is with Computer Science Department, Aberystwyth University, SY23 3FL, UK. Email: jungonghan77@gmail.com

*Corresponding authors: Qiang Zhang and Jungong Han.

cases. For example, as shown in the first row of Fig. 1, although the salient objects share similar appearances with the backgrounds, the salient objects can be accurately detected with the aid of the cross-modal complementary information captured by using linear fusion strategies (e.g., concatenation). However, for many practical cases, the linear fusion strategies may not fully capture the complementary information within RGB-D images for saliency detection. For examples, as shown in the last two rows of Fig. 1, the input RGB-D images consist of multiple salient objects with large variants. Moreover, the depth information of these salient objects is also different. In these cases, as shown in Fig. 1(e), only one of the salient objects may be highlighted by employing the linear fusion strategies. This may result from that, in these cases, the relationships between the unimodal RGB and depth information are too complicated to be captured by using linear fusion strategies. To address this issue, a Multi-modal Feature Interaction (MFI) module is designed in our proposed model to capture more complementary information from input RGB-D images by exploring more interactions between RGB and depth features.

In the proposed MFI module, a Multi-modal Bilinear Fusion (MBF) submodule is first employed to capture pairwise interactions between the unimodal RGB and depth features by employing a bilinear feature fusion strategy. Then, as suggested in [30], on the top of MBF submodule, a Multi-modal Linear Fusion (MLF) submodule is designed to preserve those highly discriminative unimodal RGB and depth features. By virtue of the proposed MFI module, more interactions between unimodal RGB and depth features can be explored. This eventually leads to effectively cross-modal complementary information extraction and better saliency detection (e.g., Fig. 1(f)).

(2) How to fully exploit the extracted cross-modal complementary information for saliency detection.

Generally, the high-level and low-level cross-modal features are complementary to each other. The high-level cross-modal features mainly contain semantic information that may help to locate the position of the salient objects, while the low-level cross-modal features contain fine details, which help to detect the boundaries of the salient objects. Therefore, existing works try to exploit the multi-level complementary information in the different levels of cross-modal features for saliency detection [25]. For example, a fluid pyramid architecture was presented to better utilize the multi-level complementary information for saliency detection in [25]. It employed many fluid connections to provide more interactions between the cross-modal features at different levels.

However, the low-level cross-modal features contain some fine details related to the salient objects as well as some fine details related to the backgrounds. As a result, the saliency information contained in the high-level cross-modal features may be degraded by those fine details of the backgrounds in the low-level cross-modal features when the cross-modal features are fused in a coarse-to-fine way for saliency detection. This has been long ignored by most existing models. To address such issues, a Saliency Prior Information Guided Fusion (SPIGF) module is presented in this paper to better exploit the cross-level complementary information. To this end, the extracted cross-modal features at different levels are first fused to predict a coarse saliency map in the proposed SPIGF module. These coarsely fused features will contain lots of saliency information. Therefore, they are then considered as saliency priors to guide the fusion of the cross-modal features at different levels. By virtue of the SPIGF module, the cross-level complementary information within the cross-modal features at different levels may be better exploited for saliency reasoning, which will further boost the SOD task.

Besides, most existing RGB-D SOD models only employ a simple convolutional layer for the final saliency prediction. Generally, some channels of the features used for the final saliency detection mainly contain information about the foregrounds (defined as foreground features), while some other channels of features mainly contain information about the backgrounds (defined as background features). However, the foreground features may also contain some disturbing information about the background regions, and vice versa. More worriedly, these disturbing information may lead to suboptimal saliency results if only a simple convolutional layer is employed for the final saliency prediction. Considering that, a Saliency Refinement and Prediction (SRP) module is presented, in which those foreground features and background features are first refined and then used to deduce the final saliency maps. By virtue of the proposed SRP module, the accuracy of saliency detection is significantly improved.

In summary, the main contributions of this work are as follows:

(1) A novel RGB-D SOD model is presented to facilitate two core subtasks in RGB-D saliency detection, i.e., multi-modal feature fusion and saliency reasoning, which achieves state-of-the-art results on several benchmark datasets.

(2) An MFI module is presented to better capture the cross-modal complementary information between the input unimodal RGB and depth images, where the linear and bilinear feature fusion strategies are engaged. This clearly differs from most of the existing models, where only the simple linear feature fusion strategies (e.g., concatenation and element-wise addition) are utilized.

(3) An SPIGF module is designed to effectively exploit the cross-level complementary information of the fused cross-modal features at different levels. Unlike most of existing models that just adopting simple concatenation to capture the cross-level features, the SPIGF module employs some saliency prior information to guide the fusion of cross-modal features at different levels.

(4) An SRP module is presented for the final saliency prediction, where the foreground features and the background features are first refined and then used to predict the final saliency maps. This is different from most existing models, where a simple convolutional layer is often employed to deduce the final saliency maps.

The rest of this paper is organized as follows. In Section II, we briefly introduce some previous works related to RGB and RGB-D SOD. In Section III, the details of the proposed method are presented. Several experiments are conducted to

validate the proposed model in Section IV. Finally, in Section V, a brief conclusion is made.

II. Related Work

A. RGB SOD

RGB SOD has been well studied in the past two decades and many RGB SOD models have been presented. Conventional models mainly employ various of hand-designed features (e.g., color, texture, local and global contrast) to detect the salient objects [15], [31], [32], [14], [33], [17], [16], [34]. For example, in [17], a Background-Driven Salient Object Detection (BD-SOD) method was designed to more comprehensively exploit the background prior for SOD by embedding the background prior into an optimization graph.

Recently, with the rapid development of deep learning, Fully Convolutional Neural Networks (FCNs) based RGB salient detection models have become the mainstream and achieved state-of-the-art results [2], [11], [12], [13]. Many of these FCN based models capture the multi-scale multi-level information for saliency detection. For example, in [13], a Saliency Detection Network (S-Net) was presented, which introduces some dense connections to effectively integrate multi-scale features and better exploit the contexts at multiple levels. In [35], a multi-scale context-aware feature extraction module was designed to capture the multi-scale context information for saliency detection by employing four parallel dilated convolutional layers with different dilation rates. Lately, some works try to utilize the boundary information to assist SOD [36]. For example, in [36], a novel boundary-aware network was presented to enhance the boundaries of salient objects by incorporating a boundary localization stream.

B. RGB-D SOD

Recently, many RGB-D SOD models have been proposed to utilize the complementary information of RGB-D images for boosting SOD. Generally, existing RGB-D SOD models can be divided into three categories: pixel-level fusion, feature-level fusion and decision-level fusion.

For pixel-level fusion based models, the RGB-D images are concatenated as an input of four channels for SOD models. For example, in [18], the input RGB-D images were directly concatenated and fed into a modified VGG-16 net for extracting cross-modal features. After that, to better exploit the complementary information of different levels, short connections were introduced to capture the multi-level cross-modal features for SOD. For feature-level fusion based models, the unimodal features are first extracted from the input RGB-D images and then fused by some specially designed cross-modal feature fusion modules. For example, in [4], a novel complementarity-aware fusion (CA-Fuse) module was designed to explicitly learn the complementary information from the paired RGB-D images by introducing some cross-modal residual functions and complementarity-aware supervisions. For decision-level fusion based models, two saliency maps are first deduced from the input RGB and depth images by employing two separated SOD models, respectively. Then, the final saliency maps are obtained from the two maps with some weighted averaging fusion strategies. For example, in [37], a deep reinforcement learning algorithm was presented to generate the weight maps for the fusion of the two saliency maps that deduced from the unimodal input images.

The proposed model in this paper belongs to feature-level fusion based ones. However, although great progresses have been made by those feature-level fusion based models, most of them still just employ some linear fusion strategies (e.g., concatenation or element-wise addition) to fuse multi-modal features, liking our previous work [30]. As discussed in Section I, this may lead to suboptimal saliency detection results. Differently, an MFI module is designed in our proposed RGB-D SOD model, where the linear and bilinear fusion strategies are jointly employed to capture the cross-modal complementary information.

C. Deep Bilinear Model

Bilinear Model [38], [39] has been widely studied in traditional machine learning due to its powerful representation ability. Recently, integrating the bilinear model into deep CNNs (Deep bilinear models), which can aggregate the pairwise feature interactions, has shown promising results on several computer vision tasks, including fine-grained classification [40], [41] and visual question answering [42], [43].

In [38], Bilinear Pooling (BP) was first introduced in CNNs for fine-grained classification, which represented an image as a pooled outer product of features derived from two CNNs. However, high-dimensional representations and high computational complexity may seriously limit its applicability in practice. To address these issue, in [41], Compact Bilinear Pooling (CBP) was presented to simplify BP by employing one of the two technologies (Random Maclaurin [44] and Tensor Sketch [45]). Then, the computational complexity of BP were further deduced by Low-rank Bilinear Pooling (LBP) [46] in the task of fine-grained classification. Meanwhile, in [43], Multimodal Factorized Bilinear (MFB) pooling was also presented to efficiently and effectively combine multimodal features, which obtained superior performance for visual question answering.

In this paper, the deep biliear model is employed to explore the pairwise interactions between unimodal RGB and depth features for capturing more cross-modal complementary information.

III. Proposed Model

As shown in Fig. 2, the proposed RGB-D salient detection network comprises four main components: (1) A Unimodal Feature Extraction (UFE) module to extract unimodal features from input RGB and depth images; (2) A Multi-modal Feature Interaction (MFI) module to capture cross-modal complementary information between the input RGB and depth images. (3) A Saliency Prior Information Guided Fusion (SPIGF) module to exploit the extracted cross-modal features for saliency prediction; (4) A Saliency Refinement and Prediction (SRP) module to deduce the final saliency maps. In the following contents, we will discuss the four components in details, respectively.

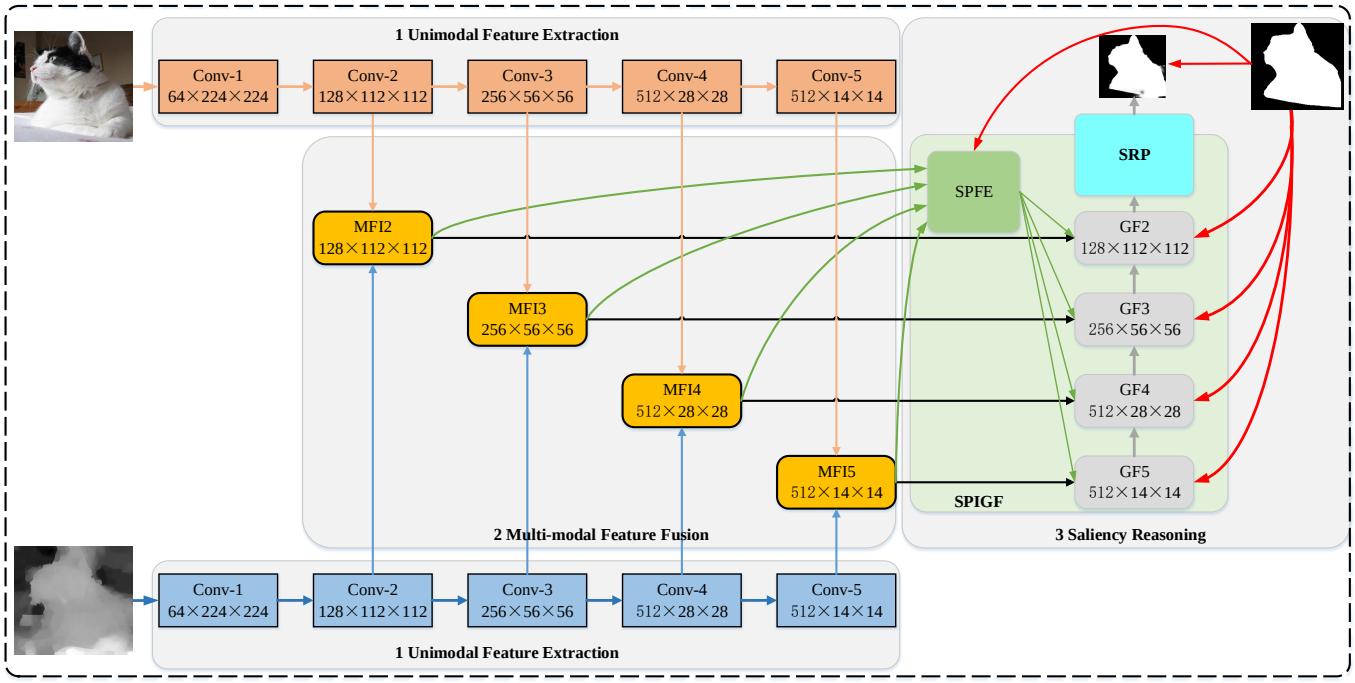


Fig. 2. Illustration of the proposed model. In the proposed model, the unimodal RGB and depth features are first extracted from the input RGB and depth images by employing two separate sub-networks, respectively. Then, in the multi-modal feature fusion, the extracted unimodal RGB and depth features are fused by using the proposed MFI module to capture cross-modal complementary information. Finally, in the saliency reasoning, the cross-level complementary information in different levels of the extracted cross-modal features are effectively exploited by using the proposed SPIGF module, and then the final saliency maps are deduced by using the proposed SRP module.

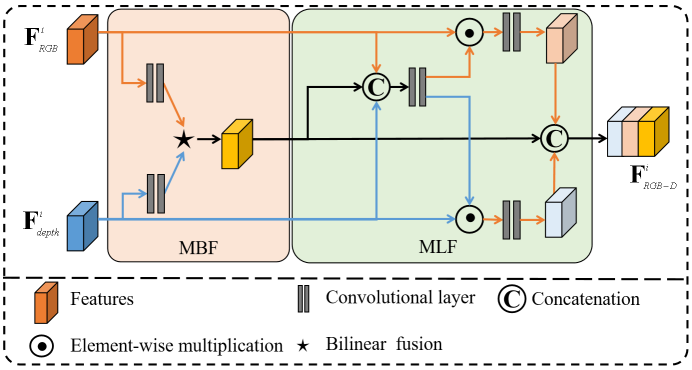


Fig. 3. Architecture of the proposed MFI module. In the proposed MFI module, the input unimodal RGB and depth features are first fused by the proposed MBF submodule to better capture the cross-modal complementary information. Then, the discriminative unimodal RGB and depth features are selected and preserved by the proposed MLF submodule.

A. UFE Module

As shown in Fig. 2, in the UFE module, the paired RGB and depth images are fed into two separate sub-networks to extract unimodal RGB and depth features of different levels, respectively. Here, in both of the two sub-networks, the VGG-16 nets [47], pre-trained on ImageNet [48], are employed as the feature extraction network for fair comparisons with previous works. Other networks, such as Res-Net [49], may also be used. Furthermore, all the fully connected layers and the last max-pooling layer are removed from the employed VGG-16 nets to preserve the spatial information. Specifically, in each sub-network, four levels of unimodal features are extracted from the input RGB or depth image, i.e., Conv 2-2 (containing 128 feature maps of size 112×112, denoted by $\mathbf{F}_m^2$), Conv 3-3 (containing 256 feature maps of size 56×56, denoted by $\mathbf{F}_m^3$), Conv 4-3 (containing 256 feature maps of size 28×28, denoted by $\mathbf{F}_m^4$), and Conv 5-3 (containing 512 feature maps of size 14×14, denoted by $\mathbf{F}_m^5$). Here $m \in \{RGB, depth\}$ denotes the RGB or depth image, respectively. The unimodal features from Conv 1-2 (i.e., $\mathbf{F}_{RGB}^1$ and $\mathbf{F}_{depth}^1$) are omitted due to its too small receptive field.

B. MFI module

RGB-D SOD is a challenging task because there are large modality gaps between the RGB images and depth images. Most existing models employ some linear fusion strategies (e.g., concatenation and element-wise addition) to fuse the unimodal RGB and depth features for capturing the cross-modal complementary information. As discussed in Section I, they may work well in most cases (e.g., the first row of Fig. 1), while they may achieve some undesirable results in some special cases, especially when the correlations between unimodal RGB and depth information are relatively complicated (e.g., the last two rows of Fig. 1). To address such an issue, a Multi-modal Feature Interaction (MFI) module is presented, where a novel multi-modal feature fusion strategy, i.e., jointly utilizing the linear and bilinear feature fusion strategies, is employed to explore more interactions between the unimodal RGB and depth features for better capturing the cross-modal complementary information.

The architecture of the proposed MFI module is shown in Fig. 3, which mainly consists of two submodules: a Multi-

modal Bilinear Fusion (MBF) submodule and a Multi-modal Linear Fusion (MLF) submodule.

1) MBF submodule: Given the unimodal features $\mathbf{F}^i_{RGB}$ and $\mathbf{F}^i_{depth} \in R^{C\times H\times W}$ at the i-th level, suppose that $X^{k,i}_R \in R^{H\times W}, k = 1, 2, ..., C$ and $X^{m,i}_D \in R^{H\times W}, m = 1, 2, ..., C$ denote the k-th and m-th channel of unimodal RGB and depth features, respectively. For element-wise addition and concatenation (i.e., the most commonly used linear fusion strategies), the cross-modal features on position (x, y) are obtained as follows:

$$\begin{aligned}\overleftarrow{\mathbf{F}}^i_{RGB-D}(x,y) &= \mathbf{F}^i_{RGB}(x,y) + \mathbf{F}^i_{depth}(x,y) = [X^{1,i}_R(x,y) \\ &+ X^{1,i}_D(x,y), X^{2,i}_R(x,y) + X^{2,i}_D(x,y), ..., X^{C,i}_R(x,y) \\ &+ X^{C,i}_D(x,y)],\end{aligned} \tag{1}$$

$$\begin{aligned}\overrightarrow{\mathbf{F}}^i_{RGB-D}(x,y) &= \mathrm{Cat}(\mathbf{F}^i_{RGB}(x,y), \mathbf{F}^i_{depth}(x,y)) = \\ &[X^{1,i}_R(x,y), X^{2,i}_R(x,y), ..., X^{C,i}_R(x,y), X^{1,i}_D(x,y), X^{2,i}_D(x,y), \\ &..., X^{C,i}_D(x,y)],\end{aligned} \tag{2}$$

where $x = 1, 2, ..., H$ and $y = 1, 2, ..., W$. $\overleftarrow{\mathbf{F}}^i_{RGB-D}(x,y) \in R^C$ denotes the features obtained by using element-wise addition operation on position (x, y). $\overrightarrow{\mathbf{F}}^i_{RGB-D}(x,y) \in R^{2C}$ denotes the features obtained by using concatenation operation on position (x, y). $\mathbf{F}^i_{RGB}(x,y)$ and $\mathbf{F}^i_{depth}(x,y)$ denote the unimodal RGB and depth features on the position (x, y), respectively. It can be seen that element-wise addition captures the interactions of two paired features and concatenation simply puts all the unimodal RGB and depth features together without capturing any interactions. As a result, these linear fusion strategies may not fully exploit the cross-modal information, especially when the correlations between unimodal RGB and depth information are relatively complicated.

To address this issue, as shown in Fig. 3, the MBF submodule is designed to explore the pairwise interactions of unimodal RGB and depth features for better capturing cross-modal complementary information. Specifically, inspired by the Bilinear Convolutional Neural Networks (B-CNNs) [40], on position (x, y), the unimodal RGB and depth features (i.e., $\mathbf{F}^i_{RGB}(x,y)$ and $\mathbf{F}^i_{depth}(x,y)$) are combined by using their outer product [40], [39], i.e.,

$$\begin{aligned}\overline{\mathbf{F}}^i_{RGB-D}(x,y) &= \mathbf{F}^i_{RGB}(x,y) \circ \mathbf{F}^i_{depth}(x,y)^T = \\ &[X^{1,i}_R(x,y) \times X^{1,i}_D(x,y), X^{1,i}_R(x,y) \times X^{2,i}_D(x,y), ..., \\ &X^{C,i}_R(x,y) \times X^{C,i}_D(x,y)].\end{aligned} \tag{3}$$

Here, $\overline{\mathbf{F}}^i_{RGB-D}(x,y) \in R^{C^2}$ denotes the fused cross-modal features on the position (x, y). $\circ$ denotes the outer product. It can be seen that, compared with those linear fusion strategies (e.g., element-wise addition shown in Eq. 1 and concatenation shown in Eq. 2), the proposed MBF submodule compares each local feature in one modality image with all of the features in the other modality image, thus obtaining a pairwise interactions of all the extracted unimodal RGB and depth features. As a result, more cross-modal complementary information may be captured from the input RGB-D images.

However, Eq. 3 also demonstrates that the bilinear feature fusion strategy has much high computational complexity. For that, following [41] and [40], the tensor sketching [45] is further employed in the proposed MBF submodule to reduce the computational complexity by aggregating low dimensional embeddings that approximate the bilinear features. [41] and [40] have proved that, compared with the original version, the improved one is simpler, faster, and equally effective.

2) MLF submodule: The proposed MBF submodule may effectively capture the cross-modal complementary information. However, as suggested in [30], besides the fused cross-modal features, the unimodal RGB and depth features may also provide useful information for saliency detection, especially when one of the input RGB and depth image has low qualities. Considering that, on the top of MBF submodule, an MLF submodule is designed to preserve those highly discriminative unimodal RGB and depth features. For that, as shown in Fig. 3, the unimodal RGB and depth features are first selected by employing a spatial-wise attention mechanism with the aid of the cross-modal features extracted by MBF submodule. Then the selected unimodal features are preserved for saliency detection.

Specifically, the extracted i-th level of cross-modal features ($\overline{\mathbf{F}}^i_{RGB-D}$), unimodal RGB and depth features ($\mathbf{F}^i_{RGB}$ and $\mathbf{F}^i_{depth}$) are first concatenated to generate two weight maps ($\mathbf{w}^i_{RGB}$ and $\mathbf{w}^i_{depth} \in R^{H\times W}$) for selecting the discriminative unimodal RGB and depth features at the i-th level by performing some convolutional layers with $\mathrm{Sigmoid}$ activation function on these concatenated features, i.e.,

$$\mathbf{w}^i_{RGB}, \mathbf{w}^i_{depth} = \sigma(\mathrm{Conv}(\mathrm{Cat}(\overline{\mathbf{F}}^i_{RGB-D}, \mathbf{F}^i_{RGB}, \mathbf{F}^i_{depth}), \mu_i)), \tag{4}$$

where $\mathrm{Conv}(*, \mu_i)$ denotes the convolutional layers with the parameters μ_i. $\sigma(*)$ denotes the $\mathrm{Sigmoid}$ activation function. $\mathrm{Cat}(*)$ denotes the concatenation operation. Given the generated weight maps, the discriminative unimodal RGB and depth features are selected by

$$\widetilde{\mathbf{F}}^i_{RGB} = \mathrm{Conv}(\mathbf{F}^i_{RGB} \odot \mathbf{w}^i_{RGB}, \beta^r_i), \tag{5}$$

$$\widetilde{\mathbf{F}}^i_{depth} = \mathrm{Conv}(\mathbf{F}^i_{depth} \odot \mathbf{w}^i_{depth}, \beta^d_i), \tag{6}$$

where $\mathrm{Conv}(*, \beta^r_i)$ and $\mathrm{Conv}(*, \beta^d_i)$ denotes two convolutional layers with the parameters β^r_i and β^d_i, respectively. $\odot$ denotes the element-wise multiplication. The final cross-modal features $\mathbf{F}^i_{RGB-D}$ in the i-th level are thus obtained by concatenating $\overline{\mathbf{F}}^i_{RGB-D}, \widetilde{\mathbf{F}}^i_{RGB}$ and $\widetilde{\mathbf{F}}^i_{depth}$, i.e.,

$$\mathbf{F}^i_{RGB-D} = \mathrm{Cat}(\overline{\mathbf{F}}^i_{RGB-D}, \widetilde{\mathbf{F}}^i_{RGB}, \widetilde{\mathbf{F}}^i_{depth}). \tag{7}$$

With the proposed MBF and MLF submodules collaborated, more interactions between unimodal RGB and depth features are explored to capture various cross-modal complementary information and, meanwhile, those high discriminative unimodal RGB and depth features are also preserved for RGB-D saliency detection.

C. SPIGF Module

The features extracted by the proposed MFI modules contain various high-level semantic information and low-level

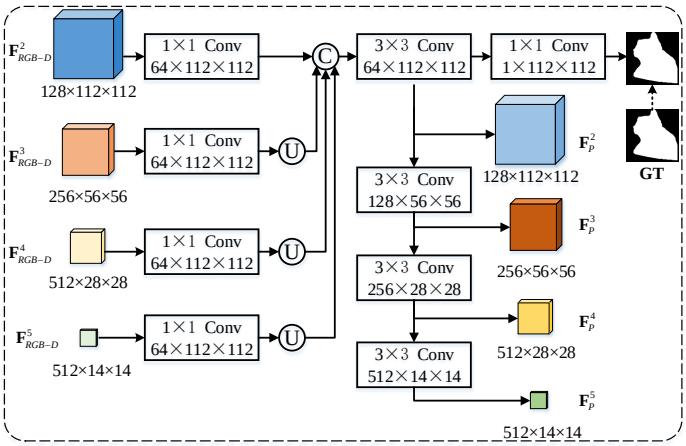


Fig. 4. Illustration of the proposed SPFE submodule. In the proposed SPFE submodule, the four levels of the extracted cross-modal features are first fused to deduce a coarse saliency map and then the saliency priors for guiding the fusion of the cross-modal features at different levels are extracted from the fused features.

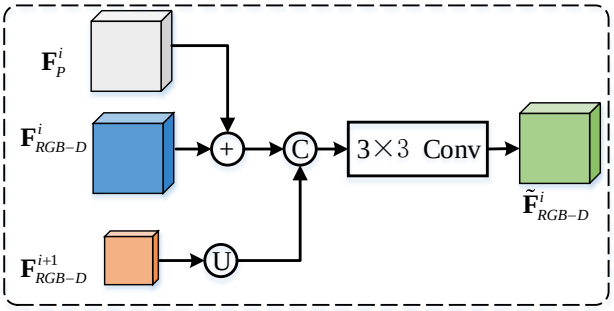


Fig. 5. Illustration of the proposed GF submodule. In the proposed GF submodule, the saliency prior features are first fused with low-level cross-modal features and then the fused features are further fused with their corresponding high-level cross-modal features to better exploit the cross-level complementary information.

visual details. Generally, for SOD task, the semantic information in deep levels may help to locate the salient objects and the visual details in shallow levels may help to identify boundaries of salient objects. However, the features in shallow levels also contain many fine details about the backgrounds which may introduce disturbing information and thus degrade the semantic information of deep layers, if the features in the shallow levels and those in the deep levels are directly fused by using some linear fusion strategies (e.g., element-wise addition and concatenation). This may lead to inaccurate or even wrong saliency prediction. To better exploit the cross-level complementary information for saliency detection, an SPIGF module is presented in this paper.

As shown in Fig. 2, the proposed SPIGF module contains two submodules, i.e., a Saliency Prior Features Extraction (SPFE) submodule and a Guided Fusion (GF) submodule. Here, the SPFE submodule is to extract the saliency prior information and the GF submodule is to fuse the cross-modal features of different levels with the aid of those saliency prior information

1) SPFE Submodule: As shown in Fig. 4, the proposed SPFE module first fuses the four levels of extracted cross-modal features to predict a coarse saliency map. Naturally, the coarse saliency map can be used as prior information to guide the fusion of cross-modal features at different levels. However, compared with the coarse saliency map, the fused features for deducing the coarse saliency map also contain lots of saliency information. Moreover, they also combine low-level fine details and high-level semantic information. Therefore, in the proposed SPFE submodule, those fused features, rather than the coarse saliency map, are used as prior information to guide the fusion of cross-modal features at different levels.

Concretely, the four levels of cross-model features (i.e., $\mathbf{F}^2_{RGB-D}$, $\mathbf{F}^3_{RGB-D}$, $\mathbf{F}^4_{RGB-D}$ and $\mathbf{F}^5_{RGB-D}$) are first reduced to 64 channels by employing four 1×1 convolutional layers, respectively. Then, the reduced features are resized to the size of $\mathbf{F}^2_{RGB-D}$ by using the bilinear operation and fused by employing a 3×3 convolutional layer. Mathematically, this process can be expressed by

$$\mathbf{F}^2_p = \mathrm{Conv}(\mathrm{Cat}(\mathrm{Conv}(\mathbf{F}^2_{RGB-D}, \theta_2), \mathrm{UP}(\mathrm{Conv}(\mathbf{F}^3_{RGB-D}, \theta_3)), \mathrm{UP}(\mathrm{Conv}(\mathbf{F}^4_{RGB-D}, \theta_4)), \mathrm{UP}(\mathrm{Conv}(\mathbf{F}^5_{RGB-D}, \theta_5))), \theta_p), \tag{8}$$

where $\mathrm{Conv}(*, \theta_p)$ denotes the 3×3 convolutional layer with parameters θ_p. $\mathrm{Conv}(*, \theta_i), i = 2, 3, 4, 5$, denotes a 1×1 convolutional layer with parameter θ_i for the reduction of the channels of the corresponding cross-modal features. $\mathrm{UP}(*)$ denotes the up-sample operation by using bilinear operation. After that, a coarse saliency maps will be deduced from the fused features, i.e.,

$$\mathbf{S}_p = \mathrm{Conv}(\mathbf{F}^2_p, \theta_{\mathbf{S}_p}), \tag{9}$$

where $\mathbf{S}_p$ denotes the saliency map deduced by $\mathbf{F}^2_p$. $\mathrm{Conv}(*, \theta_{\mathbf{S}_p})$ denotes a 1×1 convolutional layer with parameter $\theta_{\mathbf{S}_p}$.

Finally, the fused features $\mathbf{F}^2_p$ are employed as the saliency priors for the cross-modal features at the 2^{nd} level, due to the fact that $\mathbf{F}^2_p$ contains lots of saliency information and has the same size as the cross-modal features in the 2^{nd} level. Then, as shown in Fig. 4, the saliency priors for the 3^{rd}, 4^{th} and 5^{th} levels are extracted from the fused features by further employing several convolution layers, considering that the cross-modal features in the 3^{rd}, 4^{th} and 5^{th} levels have different sizes and channels with $\mathbf{F}^2_p$, i.e.,

$$\mathbf{F}^3_p = \mathrm{Conv}(\mathbf{F}^2_p, \theta^3_p), \mathbf{F}^4_p = \mathrm{Conv}(\mathbf{F}^3_p, \theta^4_p), \mathbf{F}^5_p = \mathrm{Conv}(\mathbf{F}^4_p, \theta^5_p), \tag{10}$$

where $\mathrm{Conv}(*, \theta^i_p), i = 3, 4, 5$, denote the 3×3 convolutional layers with the parameters θ^i_p. Besides, here, all of the convolutional layers set stride to 2.

2) GF Submodule: As shown in Fig. 5, the next step is to employ those extracted saliency prior features to guide the fusion of cross-modal features at different levels. Specifically, given the i-th level of cross-modal features (i.e., $\mathbf{F}^i_{RGB-D}$) and the features from the previous SPIGF module (i.e., $\widetilde{\mathbf{F}}^{i+1}_{RGB-D}$), the features at the i-th level (i.e., $\widetilde{\mathbf{F}}^i_{RGB-D}$) are obtained as follows. First, the prior information contained in the i-th level of saliency prior features $\mathbf{F}^i_p$ is introduced to $\mathbf{F}^i_{RGB-D}$ by element-wise summation. As a result of

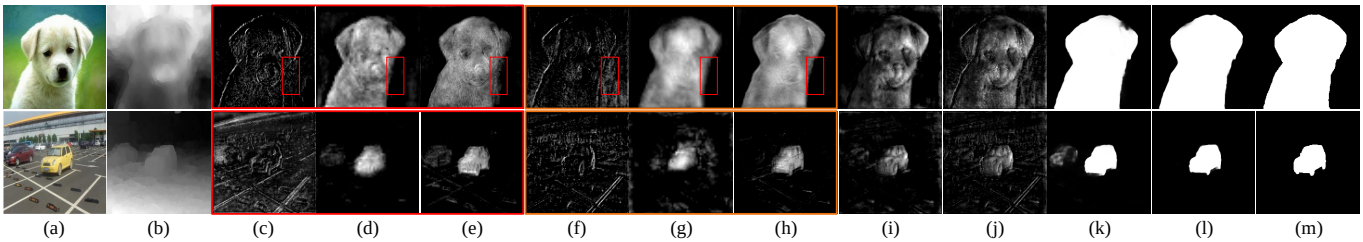


Fig. 6. Illustration of some features obtained by the proposed SPIGF module. (a) RGB images; (b) Depth images; (c)-(e) Low-level and high-level cross-modal features and corresponding fused cross-level features by using concatenation; (f)-(h) Low-level and high-level cross-modal features and corresponding fused cross-level features by using the proposed SPIGF module; (i) Saliency prior features; (j) The strengthened low-level cross-modal features; (k),(l) Saliency maps deduced from the features obtained by using concatenation and the proposed SPIGF module, respectively; (m) Ground truth.

that, some saliency information is introduced to the low-level features. This may guide the high-level cross-modal features to focus on the fine details in the saliency regions, when fusing the features of different level. After that, $\widetilde{\mathbf{F}}^{i+1}_{RGB-D}$ is then up-sampled by using the $\mathrm{bilinear}$ operation and fused with the features from the previous step to obtain more discriminative multi-level cross-modal features (i.e., $\widetilde{\mathbf{F}}^{i}_{RGB-D}$). Mathematically, this process is expressed by

$$\widetilde{\mathbf{F}}^{i}_{RGB-D} = \mathrm{Conv}(\mathrm{Cat}(\mathbf{F}^{i}_{p} + \mathbf{F}^{i}_{RGB-D}, \mathrm{UP}(\widetilde{\mathbf{F}}^{i+1}_{RGB-D})), \alpha_i), \tag{11}$$

where $\mathrm{Conv}(*, \alpha_i), i = 2, 3, 4$ denotes a 3×3 convolutional layer with the parameters α_i.

As shown in Fig. 6, compared with the fused cross-level features obtained by using concatenation, the features obtained by the proposed SPIGF module contain less disturbing information in the backgrounds. As a result, more accurate saliency results are obtained. As shown in Fig. 6(i) and (j), this may result from that the saliency prior features extracted by the proposed SPFE submodule contain lots of the saliency information and provide some guidances to better fuse the low-level and high-level cross-modal features, when introducing some saliency information to corresponding low-level features.

D. SRP Module

Given the features $\widetilde{\mathbf{F}}^{2}_{RGB-D}$ from the second level of the SPIGF module, most existing models usually employ a simple convolutional layer to deduce the final saliency map. However, as discussed in Section I, some channels of features in $\widetilde{\mathbf{F}}^{2}_{RGB-D}$ mainly contain the foreground information (defined as foreground features, e.g., Fig. 7(c)-(e), respectively), while some channels of features mainly contain the background information (defined as background features, e.g., 7(f)-(h), respectively). Furthermore, as shown in the regions marked by red boxes of Fig. 7, those foreground features inevitably contain some disturbing information related to the background regions and those background features may also contain some disturbing information related to the foreground regions, even after being processed by the proposed SPIGF module. These disturbing information may lead to suboptimal saliency results (e.g., Fig. 7(i)) if a simple convolutional layer is employed for final saliency prediction. To address this issue, an SRP module is designed in the proposed model, which first explicitly refines the foreground features and background features and then deduces the final saliency maps from the refined features.

Specifically, compared with background features (e.g., Fig. 7 (f)-(h)), most foreground features (e.g., Fig. 7 (c)-(e)) mainly have higher activations on the foreground regions, while having lower activations on the background regions, and vice versa. Naturally, the total activation values of all the foreground features are more likely to be higher than those of all the background features from the foreground regions, and vice versa. Based on that, in the SRP module, foreground features and background features are first explicitly selected. Then, the total activation values of all the foreground features and those of all the background features are computed to generate two refinement-weight maps for refining the foreground features and background features, respectively.

The structure of the proposed SRP module is shown in Fig. 8. Given the features $\widetilde{\mathbf{F}}^{2}_{RGB-D} \in R^{C \times H \times W}$, a channel-wise attention block is employed to select the foreground features and the background features in $\widetilde{\mathbf{F}}^{2}_{RGB-D}$. Specifically, a global average pool is first performed on $\widetilde{\mathbf{F}}^{2}_{RGB-D}$ to squeeze its global information into one dimension vector, and then two stacked fully connected layers with a $\mathrm{Sigmoid}$ function are employed to generate the weights $\omega \in R^{C}$ for the selection of those foreground features, i.e., :

$$\omega = \sigma(\mathrm{FC}(\mathrm{GAP}(\widetilde{\mathbf{F}}^{2}_{RGB-D}), \gamma_w)), \tag{12}$$

where $\sigma(*)$ denotes the $\mathrm{Sigmoid}$ function. $\mathrm{GAP}(*)$ denotes the global average pool and $\mathrm{FC}(*, \gamma_w)$ denotes the fully connected layer with its parameters γ_w.

Higher values in ω represent that the features of the corresponding channels are more likely to be foreground ones. Accordingly, higher values in $\mathbf{1} - \omega$ represent that the features of the corresponding channels are more likely to be background ones. Here, $\mathbf{1}$ denotes a vector of 1's with the same dimensions of ω. Based on that, the foreground and background features are thus selected by:

$$\begin{aligned} \widetilde{\mathbf{F}}_{fg} &= \widetilde{\mathbf{F}}^{2}_{RGB-D} \otimes \omega \\ \widetilde{\mathbf{F}}_{bg} &= \widetilde{\mathbf{F}}^{2}_{RGB-D} \otimes (\mathbf{1} - \omega), \end{aligned} \tag{13}$$

where $\widetilde{\mathbf{F}}_{fg}$ and $\widetilde{\mathbf{F}}_{bg} \in R^{C \times W \times H}$ denote the foreground and background features, respectively. $\otimes$ denotes the channel-wise multiplication.

The total activation maps of the foreground features and those of the background features (denoted by $\mathbf{V}_{fg}$ and $\mathbf{V}_{bg} \in$

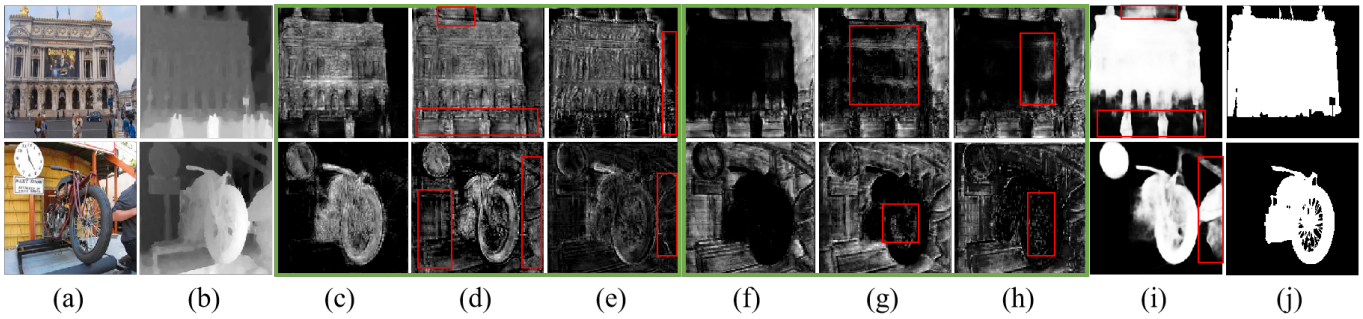


Fig. 7. Illustration of some features in $\widetilde{\mathbf{F}}^2_{RGB-D}$. (a) RGB images; (b) Depth images; (c)-(e) Foreground features; (f)-(h) Background features; (i) Saliency maps deduced by the model without SRP module. (j) Ground truth.

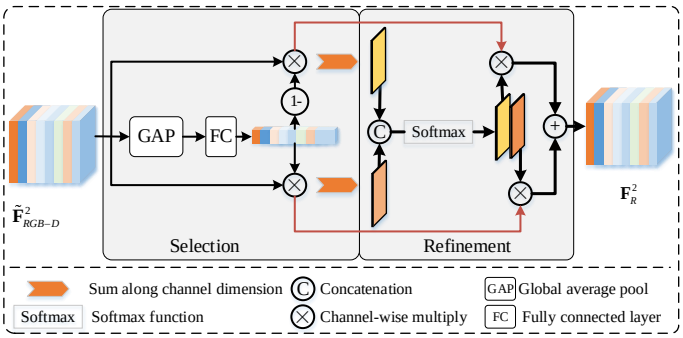


Fig. 8. Illustration of the proposed SRP module. In the proposed SPR module, the foreground features and the background features are first selected by employing a channel-wise attention block. Then, the refinement-maps are generated from the selected foreground features and the background features. Finally, the foreground features and the background features are refined by those generated refinement-maps and the final saliency maps are deduced from those refined features.

$R^{W\times H}$) are calculated by summing $\widetilde{\mathbf{F}}_{fg}$ and $\widetilde{\mathbf{F}}_{bg}$ along their channel dimensions, respectively, i.e.,

$$\mathbf{V}_{fg}=\sum_{q=1}^{C}\widetilde{\mathbf{F}}^q_{fg},\mathbf{V}_{bg}=\sum_{q=1}^{C}\widetilde{\mathbf{F}}^q_{bg}, \tag{14}$$

where $\widetilde{\mathbf{F}}^q_{fg}$ and $\widetilde{\mathbf{F}}^q_{bg}\in R^{W\times H}$ are the q-th channel of $\widetilde{\mathbf{F}}_{fg}$ and $\widetilde{\mathbf{F}}_{bg}$, respectively. Then, the refinement-weight maps (denoted by $\mathbf{R}_{fg}$ and $\mathbf{R}_{bg}\in R^{W\times H}$) for $\widetilde{\mathbf{F}}_{fg}$ and $\widetilde{\mathbf{F}}_{bg}$ are generated by normalizing $\mathbf{V}_{fg}$ and $\mathbf{V}_{bg}$ at each spatial location with $\mathrm{Softmax}$ function, i.e.,

$$\mathbf{R}_{fg},\mathbf{R}_{bg}=\mathrm{Softmax}(\mathrm{Cat}(\mathbf{V}_{fg},\mathbf{V}_{bg})), \tag{15}$$

where $\mathrm{Cat}(*)$ denotes the concatenation operation. Here, in $\mathbf{R}_{fg}$, higher values mean that the features on the corresponding positions are more likely to be foreground features, while, in $\mathbf{R}_{bg}$, higher values indicate that the features on the corresponding positions are more likely to be background features. The refined features $\mathbf{F}_R$ are thus obtained by

$$\mathbf{F}_R=\widetilde{\mathbf{F}}_{fg}\odot\mathbf{R}_{fg}+\widetilde{\mathbf{F}}_{bg}\odot\mathbf{R}_{bg}, \tag{16}$$

where $\odot$ is the element-wise multiplication. Given the refined features $\mathbf{F}_R$, the final saliency map $\mathbf{S}$ is thus obtained by performing two stacked convolutional layers on the refined features, i.e.,

$$\mathbf{S}=\mathrm{Conv}(\mathbf{F}_R,\varrho), \tag{17}$$

where $\mathrm{Conv}(*,\varrho)$ denotes the two stacked convolutional layers with the parameters ϱ.

Fig. 9 illustrates some refinement-weight maps. From Fig. 9, it can be seen that, by virtue of the proposed SRP module, the foreground and background features can be selected and refined by employing their corresponding refinement-weight maps. As a result, more accurate saliency results are deduced from these refined features.

E. Loss Function

Cross-Entropy (CE) loss and Lovász-Softmax (LS) loss between the final saliency map $\mathbf{S}$ and the ground truth $\mathbf{Y}$ are employed to train the proposed network, which can be computed by

$$\zeta_1=\mathrm{CE}(\mathbf{S},\mathbf{Y})+\mathrm{LS}(\mathbf{S},\mathbf{Y}), \tag{18}$$

where CE loss is widely used for saliency detection and is expressed by:

$$\mathrm{CE}(\mathbf{S},\mathbf{Y})=\mathbf{Y}\log(\mathbf{S})+(1-\mathbf{Y})\log(1-\mathbf{S}). \tag{19}$$

LS loss is employed to optimize the mean intersection-over-union loss, which helps to obtain fine boundaries. More details about the LS loss can bee seen in [50].

Furthermore, the intermediate supervisions as in [11], [30] are also performed on the proposed MFI modules to better capture the cross-level complementary information. Mathematically, the loss is expressed by

$$\zeta_2=\sum_{i=2}^{5}\left(\mathrm{CE}\left(\mathbf{Y}_i,\mathbf{S}_i\right)+\mathrm{LS}\left(\mathbf{Y}_i,\mathbf{S}_i\right)\right), \tag{20}$$

where $\mathbf{S}_i$ is the saliency map deduced by the output features of the MFI module in the i-th level through a 1×1 convolution layer with the $\mathrm{Sigmoid}$ function. $\mathbf{Y}_i$ is the corresponding ground truth at the i-th level, which is sampled from $\mathbf{Y}$ with the size of $\mathbf{S}_i$.

Besides, it should be noticed that the refinement-weight map $\mathbf{R}_{fg}$ generated by the SRP module should also be consistent

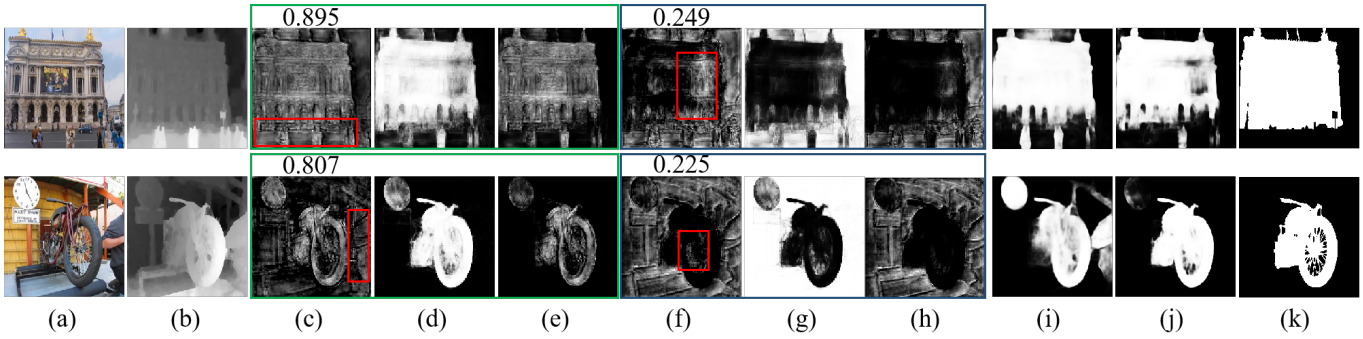


Fig. 9. Illustration of some features and corresponding weights of the proposed SRP module. (a) RGB images; (b) Depth images; (c) Foreground features and corresponding weights; (d) Refinement-weight maps for those foreground features; (e) The refined foreground features; (f) Background features and corresponding weights; (g) Refinement-weight maps for those background features; (h) The refined background features; (i) Saliency maps deduced by using a 3×3 convolutional layer with a $\mathrm{Sigmoid}$ function; (j) Saliency maps deduced by the proposed SRP module; (k) Ground truth.

with the final saliency maps. Therefore, $\mathbf{R}_{fg}$ is also supervised by the ground truth, i.e.,

$$\zeta_3 = \mathrm{CE}\left(\mathbf{R}_{fg}, \mathbf{S}\right). \tag{21}$$

Therefore, the total loss for training the proposed model is expressed by

$$\zeta = \zeta_1 + \zeta_2 + \zeta_3. \tag{22}$$

IV. Experiments

A. Implementation

The proposed model is implemented by employing the PyTorch repository [51]. An NVIDIA 1080Ti GPU is used to train and test the proposed model. The parameters of the VGG-16 nets employed in the UFE module are initialized by using the pre-trained model on ImageNet [48], while other parameters are randomly initialized by using the Xavier initialization [52]. The batch size is set to 4. The whole proposed model is trained end-to-end by using the SGD optimizer with the weight decay of 0.0005 and the initial learning rate of 0.001. Meanwhile, the learning rate is decreased by a factor of 0.8 for every 20 epochs. The input images are resized to 224×224 before being processed by using the $\mathrm{bilinear}$ operation. Meanwhile, the random flipping and random cropping are also employed for data augmentation.

B. Datasets

To evaluate the performance of our proposed network, we conduct evaluations on four public RGB-D SOD benchmarks: NJU2000 [27], NLPR [53], STEREO [54] and SIP [28]. The first three datasets are widely used to evaluate the performance of RGB-D SOD models. NJU2000 contains 1985 RGB-D image pairs collected from the internet, 3D movies and photographs. NLPR contains 1000 RGB-D image pairs under different scenarios captured by Kinect. STEREO contains 797 RGB-D images pairs.

To obtain more accurate results, K-fold cross validation is employed to train the proposed model. Concretely, the dataset is first shuffled randomly and then divided into K groups equally. Then $K-1$ groups of the RGB-D image paris are selected as the training set and the rest group of the RGB-D image pairs are employed as the testing set. In our experiments, K is set to 4 for NJU2000 and is set to 3 for NLPR. STEREO and SIP are directly tested by the models trained on NJU2000.

C. Evaluation Metrics

To evaluate our network, five popular criteria are used for performance evaluation, i.e., F-measure (F_β), $Precision$ and $Recall$ curve (PR curve), Mean Absolute Error (MAE), S-measure (S_λ) and E-measure (E_γ) [55]

F-measure (F_β) is an overall performance measurement and comprehensively considers both $Precision$ and $Recall$ by computing the weighted harmonic mean, i.e.,

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}, \tag{23}$$

where β^2 is set to 0.3, as suggested in [27]. $Precision$ and $Recall$ (PR) are standard metrics for evaluating saliency performance, which are calculated based on the binarized saliency map and the ground truth, i.e.,

$$Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN}, \tag{24}$$

where TP, TN, FP and FN denote the True-Positive, True-Negative, False-Positive, and False-Negative, respectively

MAE denotes the average errors between the saliency map $\mathbf{S}$ and the ground truth $\mathbf{Y}$, which is computed by

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |\mathbf{S}(x,y) - \mathbf{Y}(x,y)|, \tag{25}$$

where W and H are the width and height of the saliency map (or ground truth), respectively.

S-measure (S_λ) evaluates the spatial structure similarities between the saliency map $\mathbf{S}$ and the ground truth $\mathbf{Y}$, which is formulated by

$$\mathrm{S}_\lambda = \alpha * S_o + (1-\alpha) * S_r, \tag{26}$$

where $\alpha \in [0,1]$ is the balance parameter and is set to 0.5 as default. More details are seen in [56].

E-measure (E_γ) [43] combines local pixel values with the image-level mean value in one term, thus jointly capturing

TABLE I
QUANTITATIVE RESULTS OF OUR PROPOSED METHOD WITH DIFFERENT SETTINGS.

Methods	MAE	F_β	S_λ	$E\gamma$
Baseline	0.059	0.841	0.887	0.900
+MLF	0.056	0.849	0.887	0.903
+MBF	0.050	0.861	0.893	0.912
+MFI	0.046	0.876	0.898	0.921
+SPIGF	0.048	0.876	0.895	0.918
+SRP	0.050	0.865	0.889	0.904
+MFI+SPIGF	0.041	0.884	0.903	0.929
+MFI+SRP	0.043	0.885	0.903	0.927
+MFI+SPIGF+SRP	0.038	0.895	0.907	0.936

image-level statistics and local pixel matching information. It is computed by:asure (E_γ) [43] combines local pixel values with the image-level mean value in one term, thus jointly capturing image-level statistics and local pixel matching information. It is computed by:

$$E_\gamma = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \phi_{FM}(x, y), \tag{27}$$

where W and H are the width and height of saliency maps. $\phi_{FM}(*)$ is the enhanced alignment matrix whose details are in [55].

D. Ablation Analysis

To validate the effectiveness of different components in the proposed method, several experiments on the NJU2000 datasets are conducted. The proposed MFI, SPIGF and SRP modules are first removed from the proposed model and replaced by their corresponding simpler ways for saliency detection as the Baseline model. Specifically, the MFI module is replaced by element-wise concatenation to fuse multi-modal features and the SRP module is replaced by a simple 3×3 convolutional layer with a $\mathrm{Sigmoid}$ function for final saliency prediction. Then, as shown in Table I, various versions of the proposed model are designed for comparing the performance of the proposed method with different experimental settings.

As shown in Table I, compared with Baseline, the MLF submodule can boost the performance of SOD. This may due to the fact that the employed spatial attention can select those discriminative unimodal features. Furthermore, the proposed MBF submodule can significantly improve the performance of saliency detection. This may result from that MBF submodule captures more cross-modal complementary information than linear fusion strategies by exploring pairwise interactions between unimodal RGB and depth features. Meanwhile, Baseline+MFI can further improve the accuracy of SOD task. This indicates that, on the top of MBF submodule, the MLF submodule can further boost the performance of SOD by preserving those unimodal RGB and depth information.

The SPIGF module (i.e., Baseline+SPIGF) may also significantly boost the SOD task. This indicates that the SPIGF module can better exploit the extracted cross-modal information by using saliency priors to guide the fusion of multi-level cross-modal features. Similarly, the proposed SRP module (i.e., Baseline+SRP) may also boost the SOD task. This may result from that, compared with a simple convolutional layer, the proposed SRP module can better exploit the extracted cross-modal information for saliency detection. Furthermore, the saliency detection results may be significantly improved by jointly employing the proposed MFI and SRP modules or the proposed MFI and SPIGF modules (i.e., Baseline+MFI+SRP and Baseline+MFI+SPIGF). Finally, Baseline+MFI+SPIGF+SRP (i.e., the proposed model) obtains the best performance. This indicates that, with the collaboration of the proposed MFI, SPIGF and SRP modules, cross-modal complementary information is effectively captured and thoroughly exploited to boost the performance of SOD.

E. Comparison to State-of-the-Art Models

The proposed approach is compared with 11 state-of-the-art (SOTA) methods, including DF [29], MMCI [22], CTMF [5], PCA [4], TSAA [23], AF [57], CPFP [25], D3Net [28], ICNet [58], ASIFN [59], JCUF[30], DMRA [61], A2DE [60], DANet [62] and CMWN [63]. For fair comparisons, the saliency maps of these SOTA models are obtained from their authors or the deployment codes provided by their authors. The quantitative results of these models are shown in Table II and Fig. 10. The visualization results are shown in Fig. 11.

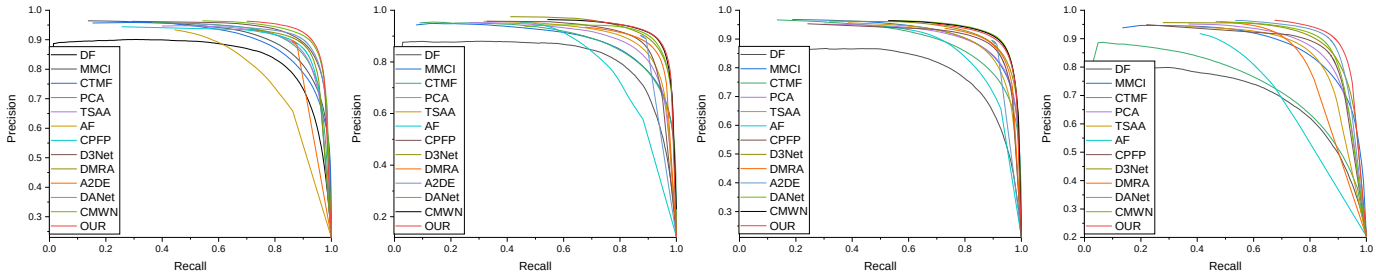
As shown in Table II and Fig. 10, for NJU2000 and SIP, the proposed model outperforms all the SOTA methods considering the five performance metrics. For NLRP and STEREO, the proposed model achieves competitive results. Fig. 11 further illustrates the performance of different models. The first two rows of Fig. 11 show some simple scenes, where all of the SOTA models can identify the salient objects. In the third and fourth rows of Fig. 11, the salient objects share similar appearances or contrast with the backgrounds. Most SOTA models fail to accurately detect salient objects, while the proposed model obtains more complete results. This indicates that the cross-modal complementary information may be better exploited for saliency detection by the proposed model. Some large objects with large internal variations are shown in the fifth and sixth rows of Fig. 11. These variations are caused by different colors or the change of appearances with their depths. For such objects, most SOTA models only detect partial salient regions due to the fact that cross-modal features extracted by using linear fusion strategies may not capture the complex correlations between RGB and depth images. While, the proposed model obtains more accurate and complete salient objects . This may owe to the proposed MFI module, which can better capture those complex interactions between RGB and depth images. Besides, as shown in the last three rows of Fig. 11, the proposed model still performs well under some complex scenes.

In addition, the inference time of nine methods, including MMCI [22], PCA [4], TSAA [23], AF [57], CPFP [25], D3Net [28], DMRA [61], DANet [62] and our proposed model, are also provided in Fig. 12. It can be seen that our model is comparable to other models in terms of inference efficiency.

TABLE II

QUANTITATIVE RESULTS BY USING DIFFERENT METHODS. MEAN F-MEASURE (F_β), S-MEASURE (S_λ), E-MEASURE ($E\gamma$) AND MEAN ABSOLUTE ERROR (MAE) ARE EMPLOYED FOR COMPARISONS. FOR F_β, S_λ AND $E\gamma$, HIGHER VALUES ARE DESIRABLE AND FOR MAE, LOWER VALUES ARE DESIRABLE.

Datasets	Metrics	DF [29]	MMCI [22]	CTMF [5]	PCA [4]	TSAA [23]	AF [57]	CPFP [25]	D3Net [28]	ICNet [58]	ASIFN [59]	JCUF [30]	A2DE [60]	DMRA [61]	DANet [62]	CMWN [63]	our
NJU2000 [27]	MAE	0.141	0.078	0.084	0.059	0.060	0.099	0.053	0.051	0.052	0.047	0.041	0.050	0.050	0.047	0.045	**0.038**
	F_β	0.649	0.793	0.778	0.839	0.841	0.765	0.850	0.860	0.869	0.877	0.881	0.869	0.873	0.874	0.881	**0.895**
	S_λ	0.762	0.858	0.849	0.876	0.879	0.773	0.895	0.895	0.894	0.888	0.901	0.868	0.885	0.897	0.902	**0.907**
	$E\gamma$	0.696	0.851	0.846	0.895	0.895	0.826	0.910	0.912	0.913	0.921	0.926	0.912	0.919	0.920	0.928	**0.936**
NLPR [53]	MAE	0.084	0.059	0.056	0.043	0.041	0.058	0.035	0.033	0.029	0.030	0.030	0.029	0.031	0.030	0.029	**0.028**
	F_β	0.664	0.736	0.740	0.802	0.819	0.755	0.840	0.852	0.884	0.874	0.885	0.875	0.864	0.871	0.877	**0.887**
	S_λ	0.801	0.855	0.859	0.873	0.886	0.799	0.888	0.905	**0.926**	0.906	0.900	0.895	0.898	0.908	0.917	0.909
	$E\gamma$	0.754	0.841	0.840	0.887	0.901	0.850	0.917	0.923	0.939	0.937	0.932	0.940	0.939	0.933	0.938	**0.940**
STEREO [54]	MAE	0.140	0.067	0.086	0.063	0.059	0.075	0.051	0.048	0.044	0.049	0.045	0.043	0.048	0.047	0.043	**0.041**
	F_β	0.616	0.812	0.758	0.818	0.827	0.806	0.841	0.844	0.869	0.864	0.870	**0.879**	0.867	0.857	0.872	0.873
	S_λ	0.757	0.872	0.848	0.874	0.871	0.824	0.879	0.890	**0.902**	0.868	0.895	0.884	0.885	0.892	0.900	0.900
	$E\gamma$	0.691	0.873	0.841	0.887	0.893	0.872	0.912	0.908	0.925	0.907	0.924	0.930	**0.930**	0.914	0.925	0.926
SIP [28]	MAE	0.185	0.086	0.139	0.070	0.075	0.117	0.063	0.062	-	-	0.056	-	0.085	0.053	0.062	**0.052**
	F_β	0.464	0.770	0.607	0.814	0.803	0.701	0.820	0.832	-	-	0.854	-	0.819	0.863	0.850	**0.863**
	S_λ	0.652	0.832	0.715	0.842	0.834	0.720	0.850	0.864	-	-	0.873	-	0.805	0.876	0.867	**0.877**
	$E\gamma$	0.564	0.844	0.704	0.878	0.870	0.792	0.893	0.893	-	-	0.904	-	0.843	0.910	0.905	**0.911**



(a) NJU2000 [27] (b) NLPR [53] (c) STEREO [54] (d) SIP [28]

Fig. 10. PR curves of different salient object detection methods on different datasets.

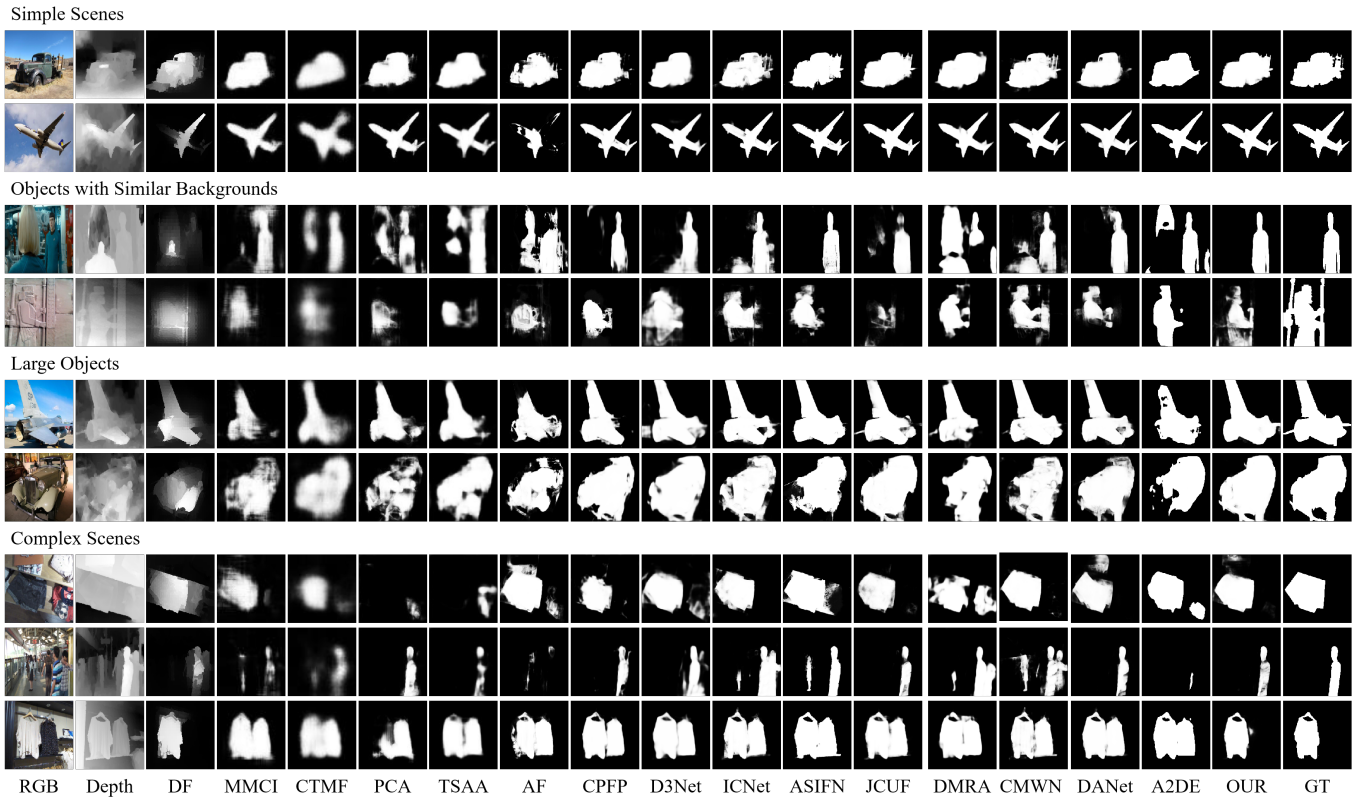


Fig. 11. Visualization results of different models. As shown in the first two rows, most existing models work well in simple scenes. However, as shown in the rest of rows, most existing works may fail to detect those salient objects under some challenging cases (e.g., objects with similar backgrounds, large objects and complex backgrounds), while the proposed model can still obtain good saliency results.

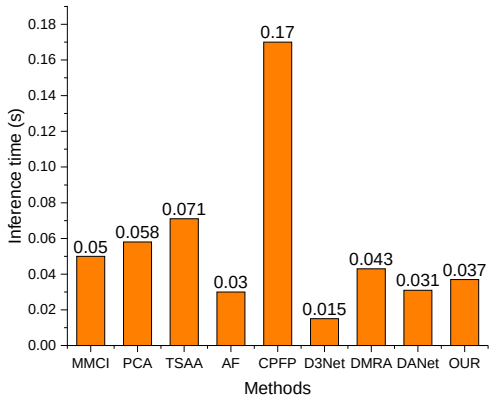


Fig. 12. The inference time of different methods.

It reveals that the running speed of our proposed method is not slow.

V. Conclusion

In this paper, a novel RGB-D SOD model is presented to facilitate two core subtasks in RGB-D saliency detection, i.e., multi-modal feature fusion and saliency reasoning. First, the proposed MFI module presents a novel multi-modal feature fusion strategy to effectively capture the cross-modal complementary information by jointly employing some linear and bilinear feature fusion strategies. By virtue of the MFI module, more interactions between unimodal RGB and depth features are explored and the cross-modal complementary information is more effectively exploited. Then, an SPIGF module and an SRP module are proposed to better exploit the extracted cross-modal information for saliency reasoning. The SPIGF module employs the saliency prior information to guide the fusion of cross-modal features within different levels. By virtue of the SPIGF module, the cross-level complementary information within different levels of cross-modal features is effectively exploited. Instead of employing a simple convolutional layer, the SRP module first refines the foreground and background features in the cross-modal features and then deduce the final saliency maps by using the refined features. With the collaboration of the MFI , SPIGF and SRP modules, the proposed model achieves the new state-of-the-art results on several benchmark datasets.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grant No. 61773301 and 61876140, and the China Postdoctoral Support Scheme for Innovative Talents under Grant No. BX20180236.

References

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[2] Y. Liu, J. Han, Q. Zhang, and C. Shan, "Deep salient object detection with contextual information guidance," *IEEE Transactions on Image Processing*, vol. 29, pp. 360–374, 2019.

[3] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3183–3192.

[4] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3051–3060.

[5] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion." *IEEE Transactions on Cybernetics*, vol. 48, pp. 3171–3183, 2018.

[6] Q. Zhang, T. Xiao, N. Huang, D. Zhang, and J. Han, "Revisiting feature fusion for rgb-t salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[7] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "Rgb-t salient object detection via fusing multi-level cnn features," *IEEE Transactions on Image Processing*, vol. 29, pp. 3321–3335, 2019.

[8] B. Lei, E.-L. Tan, S. Chen, D. Ni, and T. Wang, "Saliency-driven image classification method based on histogram mining and image score," *Pattern Recognition*, vol. 48, no. 8, pp. 2567–2580, 2015.

[9] C. Ma, Z. Miao, X. Zhang, and M. Li, "A saliency prior context model for real-time object tracking," *IEEE Transactions on Multimedia*, vol. 19, no. 11, pp. 2415–2424, 2017.

[10] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 20–33, 2017.

[11] Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr, "Deeply supervised salient object detection with short connections," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 4, pp. 815–828, 2019.

[12] Y. Tang and X. Wu, "Salient object detection using cascaded convolutional neural networks and adversarial learning," *IEEE Transactions on Multimedia*, vol. 21, pp. 2237–2247, 2019.

[13] H. Xiao, J. Feng, Y. Wei, M. Zhang, and S. Yan, "Deep salient object detection with dense connections and distraction diagnosis," *IEEE Transactions on Multimedia*, vol. 20, pp. 3239–3251, 2018.

[14] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Computational Visual Media*, vol. 5, pp. 117–150, 2014.

[15] Y. Liu, J. Han, Q. Zhang, and L. Wang, "Salient object detection via two-stage graphs," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, pp. 1023–1037, 2019.

[16] Y. Zhou, A. Mao, S. Huo, J. Lei, and S. Kung, "Salient object detection via fuzzy theory and object-level enhancement," *IEEE Transactions on Multimedia*, vol. 21, pp. 74–85, 2019.

[17] Z. Wang, D. Xiang, S. Hou, and F. Wu, "Background-driven salient object detection," *IEEE Transactions on Multimedia*, vol. 19, pp. 750–762, 2017.

[18] P. Huang, C. Shen, and H. Hsiao, "RGB-D salient object detection using spatially coherent deep learning framework," in *Proceedings of the IEEE International Conference on Digital Signal Processing*, 2018, pp. 1–5.

[19] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," *IEEE Transactions on Cybernetics*, vol. 43, pp. 1318–1334, 2013.

[20] H. Chen, Y. Li, and D. Su, "Attention-aware cross-modal cross-level fusion network for RGB-D salient object detection," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018, pp. 6821–6826.

[21] C. Zhu, X. Cai, K. Huang, T. H. Li, and G. Li, "PDNet: Prior-model guided depth-enhanced network for salient object detection," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2018, pp. 199–204.

[22] H. Chen, Y. Li, and D. Su, "Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection," *Pattern Recognition*, vol. 86, pp. 376–385, 2019.

[23] H. Chen and Y. Li, "Three-stream attention-aware network for RGB-D salient object detection," *IEEE Transactions on Image Processing*, vol. 28, pp. 2825–2835, 2019.

[24] R. Cong, J. Lei, H. Fu, J. Hou, Q. Huang, and S. Kwong, "Going from RGB to RGB-D saliency: A depth-guided transformation model." *IEEE Transactions on Cybernetics*, 2019.

[25] J. Zhao, Y. Cao, D. Fan, M. Cheng, X. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for RGB-D salient object detection,"

in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3927–3936.

[26] Z. Liu, S. S. Shi, Q. Duan, W. Zhang, and P. Zhao, "Salient object detection for RGB-D image by single stream recurrent convolution neural network," *Neurocomputing*, vol. 363, pp. 46–57, 2019.

[27] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGB-D salient object detection: A benchmark and algorithms," in *Proceedings of the European Conference on Computer Vision*, 2014, pp. 92–109.

[28] D. Fan, Z. Lin, Z. Zhang, M. Zhu, and M. Cheng, "Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[29] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "RGB-D salient object detection via deep fusion," *IEEE Transactions on Image Processing*, vol. 26, pp. 2274–2285, 2017.

[30] N. Huang, Y. Liu, Q. Zhang, and J. Han, "Joint cross-modal and uni-modal features for RGB-D salient object detection," *IEEE Transactions on Multimedia*, 2020, doi:10.1109/TMM.2020.3011327.

[31] L. Yi, Q. Zhang, J. Han, and L. Wang, "Salient object detection employing robust sparse representation and local consistency," *Image Vision Computing*, vol. 69, pp. 155–167, 2017.

[32] Q. Zhang, Y. Liu, S. Zhu, and J. Han, "Salient object detection based on super-pixel clustering and unified low-rank representation," *Computer Vision and Image Understanding*, vol. 161, pp. 51–64, 2017.

[33] M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2014.

[34] J. Lei, B. Wang, Y. Fang, W. Lin, P. L. Callet, N. Ling, and C. Hou, "A universal framework for salient object detection," *IEEE Transactions on Multimedia*, vol. 18, pp. 1783–1795, 2016.

[35] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, "A bi-directional message passing model for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1741–1750.

[36] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7479–7489.

[37] X. Wang, T. Sun, R. Yang, C. Li, B. Luo, and J. Tang, "Quality-aware multimodal saliency detection via deep reinforcement learning," *arXiv preprint arXiv:1811.10763*, 2018.

[38] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1449–1457.

[39] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Computation*, vol. 12, no. 6, pp. 1247–1283, 2000.

[40] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1449–1457.

[41] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 317–326.

[42] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 12, pp. 5947–5959, 2018.

[43] M. Lao, Y. Guo, H. Wang, and X. Zhang, "Multimodal local perception bilinear pooling for visual question answering," *IEEE Access*, vol. 6, pp. 57 923–57 932, 2018.

[44] P. Kar and H. Karnick, "Random feature maps for dot product kernels," in *Artificial Intelligence and Statistics*, 2012, pp. 583–591.

[45] N. Pham and R. Pagh, "Fast and scalable polynomial kernels via explicit feature maps," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013, pp. 239–247.

[46] S. Kong and C. Fowlkes, "Low-rank bilinear pooling for fine-grained classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 365–374.

[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations*, 2015, pp. 1–14.

[48] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 770–778.

[50] M. Berman, A. Rannen Triki, and M. B. Blaschko, "The lovász-softmax loss: a tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4413–4421.

[51] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8026–8037.

[52] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[53] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *Proceedings of the IEEE International Conference on Image Processing*, 2014, pp. 1115–1119.

[54] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 454–461, 2012.

[55] D. Fan, C. Gong, Y. Cao, B. Ren, M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *International Joint Conference on Artificial Intelligence*, 2018, pp. 698—-704.

[56] D. Fan, M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4548–4557.

[57] N. Wang and X. Gong, "Adaptive fusion for RGB-D salient object detection," *IEEE Access*, vol. 7, pp. 55 277–55 284, 2019.

[58] G. Li, Z. Liu, and H. Ling, "ICNet: Information conversion network for RGB-D based salient object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 4873–4884, 2020.

[59] C. Li, R. Cong, S. Kwong, J. Hou, H. Fu, G. Zhu, D. Zhang, and Q. Huang, "ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection," *IEEE Transactions on Cybernetics*, 2020.

[60] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu, "A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9060–9069.

[61] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7254–7263.

[62] X. Zhao, L. Zhang, Y. Pang, H. Lu, and L. Zhang, "A single stream network for robust and real-time RGB-D salient object detection," in *European Conference on Computer Vision*, 2020, pp. 646–662.

[63] G. Li, Z. Liu, L. Ye, Y. Wang, and H. Ling, "Cross-modal weighting network for rgb-d salient object detection," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 665–681.

Nianchang Huang received the B. S. degree and the M. S. degree from Qingdao University of Science and Technology, Qingdao, China, in 2015 and 2018. He is currently pursuing the Ph.D. degree in School of Mechano-Electronic Engineering, Xidian University, China. His research interests include deep learning and multimodal image processing in computer vision.

Yang Yang received his B. S. degree from Chang'an University, Xi'an, China, in 2019. He is currently pursuing the Ph.D. degree in School of Mechano-Electronic Engineering, Xidian University, China. His current research interests include multimodal image processing and deep learning.

Dingwen Zhang received his Ph.D. degree from the Northwestern Polytechnical University, Xi'an, China, in 2018. He is currently an associate professor in School of Machine-Electronical Engineering, Xidian University. From 2015 to 2017, he was a visiting scholar at the Robotic Institute, Carnegie Mellon University. His research interests include computer vision and multimedia processing, especially on saliency detection, video object segmentation, and weakly supervised learning.

Qiang Zhang received the B.S. degree in automatic control, the M.S. degree in pattern recognition and intelligent systems, and the Ph.D. degree in circuit and system from Xidian University, China, in 2001,2004, and 2008, respectively. He was a Visiting Scholar with the Center for Intelligent Machines, McGill University, Canada. He is currently a professor with the Automatic Control Department, Xidian University, China. His current research interests include image processing, pattern recognition.

Jungong Han is currently a Full Professor and Chair in Computer Science at Aberystwyth University, UK. His research interests span the fields of video analysis, computer vision and applied machine learning. He has published over 180 papers, including 40+ IEEE Trans and 40+ A* conference papers.